THE KATZ RESEARCH INSTITUTE
FOR HEBREW LITERATURE
TEL-AVIV UNIVERSITY

THE ASSOCIATION FOR LITERARY
AND LINGUISTIC COMPUTING
(A L L C)

OFFPRINT FROM:

Proceedings of

The International Conference on

LITERARY AND LINGUISTIC COMPUTING

Edited by

ZVI MALACHI, Ph.D.

ISRAEL, April 22 - 27, 1979

THE HISTORICAL DICTIONARY OF THE HEBREW LANGUAGE

Prof. Ze'ev Ben-Ḥayyim

President of the Academy of the Hebrew Language, Jerusalem

Ladies and Gentlemen,

In the name of the Academy of the Hebrew Language, I take the opportunity, offered by the opening of this session, to welcome the delegates to the conference with the biblical greeting, "Bərukhim habba'im" - "Blessed be they that come". A single goal unites all the delegates, representing as they do different countries, nations and tongues - that is, the examination of new methods in philological and linguistic research and ways of utilizing modern electronic technology to achieve higher degrees of efficiency and thoroughness in the study of the languages and the literatures favored by each of us. The very exchange of knowledge and opinions ensures both the success of the conference and the results which all of us expect of it.

The Academy of the Hebrew Language is the oldest Hebrew academic institution in our country; it came into being along with the movement for the revival of Hebrew speech in order to assist in transforming written Hebrew into a spoken, vital language permeating all walks of life.

The Hebrew language has an unbroken literary tradition that goes back three thousand two hundred years and more, but over a long period of time, ever since the loss of political independence in ancient days, it gradually

became reduced to serve as a vehicle of literature and liturgy alone. The movement for the revival of spoken Hebrew, which had its beginnings over a hundred years ago, found itself in need of a body which would guide speakers and writers, and decide between various, occasionally opposing, tendencies which had crystallized in the written tongue over so many generations and in so many linguistically varied centres.

Thus, the very founding of the institution had a decidedly pragmatic and national aim; nonetheless, even in those early days the founding fathers were already dreaming of Hebrew language research projects designed to reveal fully the past of the language, concealed as it then was in ancient manuscripts on the one hand, and in still-extant popular oral traditions on the other. The founders of this institution first raised the possibility of compiling a Hebrew dictionary encompassing all periods, and Ben-Yehuda, its president, was the man who took it upon himself to realize this dream. Ben-Yehuda's dictionary, also known by its Latin name Thesaurus Totius Hebraitatis, which took seventy years to compile (the work having been concluded only twenty years ago), can only be regarded as a first step, when we consider the vast amount of literary material being constantly revealed, to this very day, and I need mention only the Dead Sea Scrolls.

Shortly after the establishment of the State of Israel, this body became state-endowed and adopted the name The Academy of the Hebrew Language, and Hebrew language research became one of its central functions, no less significant than its practical and normative functions. Some of the problems

involved in the compilation of the Historical Dictionary, the most important of the Academy's lexicographical projects, will be demonstrated today.

By fortuitous circumstance, about twenty-five years ago, when the Academy began to plan its Historical Dictionary and to establish the project on a new basis, the first steps were taken abroad to harness computers to lexicographical endeavour. It is with great appreciation that I mention here the name of Professor Roberto Busa, the pioneer in this field who first suggested the idea which we speedily adopted - despite some hesitation and some doubt on the part of veteran lexicographers - and to which we have steadfastly adhered, while attempting to clarify the ways in which these powerful electronic instruments might be employed in the field of Hebrew linguistics. For it is to be realized, that Hebrew is a Semitic language, and as such, its morphological structure differs from those of Indo-European languages. It was thus necessary to solve problems to which no solutions had previously been sought. And so, because of these factors, the Historical Dictionary had to find its own way towards solving the methodological problems involved in the computing of Hebrew lexicographical work.

I should like to put before you something of the assumptions and goals underlying this project.

(a) Our Historical Dictionary is designed to remain current. In other words, its comprehensive aim is not just the compilation of its first edition, regardless of the time factor involved. The material accumulating has been arranged to serve as a documentation archive for the Hebrew language, which will in its turn make the compilation of the Dictionary possible. In time, these archives will be placed at the disposal of researchers in the different fields of Hebrew language and literature, such as belles lettres, philosophy or Jewish law. At present, we have already completed the preparation of all material relating to texts dating over a five-hundred-year period from 200 B.C.E., which is now assembled in 81 volumes.

(b) The material is arranged in two card-files: (1) an alphabetical card-file, containing each dictionary item together with chronologically-arranged excerpts from the sources; (2) a lexicological card-file containing the very same items arranged by structure-patterns. This card-file is of course needed for our lexicographical work, but it is also likely to provide all the material necessary for a diachronic analysis of Hebrew morphology.

(c) Meanwhile, we have been working within three sections:

(1) The Ancient Literature Section, from 200 B.C.E. up until the eleventh century C.E. The material in this section is collated directly from manuscripts, not from printed books or critical editions.

(2) The Modern Literature Section, from 1750 C.E. to the present day. The material in this section is taken from printed texts, the earliest editions always serving as preferred bases.

(3) The Computing Section, where the material prepared by our scientific researchers is adapted to the requirements of the computer.

At present, the two sections so engaged have processed some six million quotations, in various stages of preparation. The Ancient Literature Section has, in addition, a skeleton staff working on the collation of vocalized words appearing in manuscripts antedating the thirteenth century C.E. where their vocalization signs differ from the standard, in order to be able to trace the precise development of the spelling of any given lexical item.

We are not yet able to open either a section to process the texts dating from the eleventh century to 1750, or a bibliographical section parallel to the order of items in the Dictionary. This is still our plan for the future, hopefully the not-too-distant future.

Today we shall raise several of the problems peculiar to our work.

Computer-Assisted Pre-Edition of Hebrew Manuscripts

Prof. Israel Yeivin

The Academy of the Hebrew Language, Jerusalem

In the Ancient Literature section of the Historical Dictionary of the Hebrew Language Project, we are engaged in preparing concordances of the most important - both literarily and linguistically - Hebrew language compositions. In later periods only some of the compositions will be processed, whereas in the ancient period most of those written before the year 1000 C.E. are being processed. These texts are presented according to their manuscript form, for ancient textual versions are known to have undergone many contextual and linguistic changes while being copied over and over both by hand and in print. We are interested in those texts closest in time to when they were first written. The texts used by our Project have therefore been based on manuscripts, and we have to date processed hundreds of manuscripts, some lengthy, others fragmental, some whole and well-preserved, others fragmented, torn and blurred.

Our scientific assistants do their work in two stages:

A. The Textual Stage (the stage of the sentence).

The editorial board selects the oldest and best manuscript of each work to be processed, whereupon the worker prepares the text according to this manuscript. Old manuscripts generally contain unclear letters or words, additions, omissions, corrections, et al., and there are often numerous scribal errors as well. The text of the manuscript must be processed precisely, because it provides documentation of the Hebrew language as it

was a thousand or two thousand years ago, and no alterations may be made in antequatedly spelled words or unusually used ones. Our transcription must reflect all uncertainties, corrections and other orthographic phenomena, in order that the Dictionary editors will have access to a lexical card supplying the relevant context and faithfully reflecting the manuscript version, so that there will be no need to reconsult the manuscript. We attempt, in addition, to discard at this point all material we have no need of in the second stage of our work, i.e., all material not needed for the Language archives, mainly foreign language phrases, and quotations from sources already processed by the Project. On the other hand, not all such material is to be discarded, for there are loan words originating in other tongues, which have already been integrated into Hebrew; there are phrases and concepts, originally quoted from earlier works, which have been thoroughly absorbed by the quoting composition, becoming an organic part of it. We have thus to distinguish between foreign phrases and quotations which are not part of Hebrew, and those which are. The criterion for making such a distinction is a syntactic one: if the foreign phrases or the quotations are independent sentences, we do not consider them part of the language of the source being processed and we discard them, so as not to be encumbered by them in the second stage of our work. If they are not independent sentences, but rather interlaced with Hebrew words and phrases of the source being processed, we regard them as part of the language, and as such, suitable for processing in the second Stage.

After copying the manuscript and adding all the symbols needed, it is punched into the computer. A number of galleys are prepared, and are compared again and again with the manuscript itself, until we can assume that the transcribed text is free of errors.

B. The Lemmatization Stage (The stage of the word).

At this point we have the computer edit the whole text it absorbed previously, and list each word on a separate line. Our scientific worker has to jot down the lexical entry (lemma) to which each word is attributed. This is a necessary step when dealing with Hebrew because of the many affixes a word may exhibit. Were we not to list the dictionary item, but, instead, organize alphabetically the words of the text as they appear, over 50% of the words would not be grouped under the appropriate lexical entry.

In addition, we write down the symbols required to convey necessary details of the way the word appears in the manuscript - whether it is incorrectly written, or the result of a correction, etc., and also whether it is a foreign word or a quote et al. - whenever the signs used at the earlier stage are insufficient. The symbols we use in the two stages of our work agree with one another as much as possible.

After these two stages - after the text has been pre-edited, and the lemmas listed, punched into the computer, and checked - the computer automatically edits a concordance of the text, together with indices, statistics, and lexicological and lexicographical lists as well.

I shall describe one small part of our work: the symbols we use in transcribing the text. In deciding upon the system of symbols, the following three principles were adopted:

a. The computer at our disposal(C.D.C., belonging to the Hebrew University) must be able to print them.

b. The symbols must be equally applicable to all the texts to be processed; for these were written in different periods, the manuscripts recopied in different countries, etc.

c. The symbols should be easily understood. We edit texts with the aid of a computer, and desire to get texts as similar as possible to scientifically-edited texts, prepared manually, so that even readers unaccustomed to computerized work will still be able to use them. These same symbols can also, when required, serve as computer instructions, and various word-lists can be prepared with their aid.

I do not propose to describe all the symbols we use in our work, but I do intend to demonstrate several of the most common. One of the principles our symbol system of stage one, the textual stage, is based on is that they appear in pairs, namely, they enclose the phenomenon they mark, e.g., parentheses, brackets, etc. Some of the symbols take on different forms, depending on whether they precede or follow a given feature, like the various kinds of brackets. Others look the same in both positions, such a quotation-, question- and exclamation-marks. If, however, we are able in the future to increase the number of signs our computer will print, we propose to improve the system so that each

member of any given pair of symbols will be distinctive. The computer checks the symbol balance, and informs us wherever an initial sign appears without its corresponding final symbol, or vice versa.

We shall note first those symbols relating to the reading of the text.

An unclear letter or group of letters in the manuscript, whose correct reading is uncertain, is enclosed in question-marks:

(1) Manuscript: QUESTION Pre-edition: QUESTIO?N?

(2) " THE QUESTION " T?HE QUESTI?ON

Letters or words erased in the manuscript are enclosed in parentheses:

(3) Manuscript: TO∅ Pre-edition: TO(O)

(4) " TO BE ~~BE~~ OR " TO BE (BE) OR

Letters or words added in the manuscript are enclosed in square brackets:

(5) Manuscript: OR NO^{T} TO Pre-edition: OR NO[T] TO

(6) " TO BE BE OR NOT TO " TO [BE OR NOT TO] BE

Letters or words corrected in the manuscript are enclosed in both parentheses and square brackets:

(7) Manuscript: TO BE O^{R}/F Pre-edition: O(F)[R]

(8) THAT IS ~~TO SUFFER~~ THE QUESTION THAT IS (TO SUFFER)[THE QUESTION]

These are instances of simple corrections, but the manuscripts exhibit many more comlex situations. Some corrections leave the original word unclear, or only partially clear; others leave the added word unclear; some are incomplete, with not all the required letters having been erased or added; in some cases part of the correction appears on the line, and the rest of it in the margin; sometimes a single word has been corrected to two words, or two corrected to be read as one; a final letter in one word has occasionally been moved over to serve as the initial letter in the following word; a correction is sometimes found on the sheet without it being clear which word it comes to correct, etc. Tens and even hundreds of rules have crystallized over the years of manuscript analysis in our project, determining how to denot the various kinds of corrections.

As for the signs denoting quotations and foreign phrases, we distinguish between two kinds of quote: those in which the Project has no interest (since the source of the quote has already been processed and appears in our archives), such a Biblical quotes, which are met with thousands of times in the literature we deal with. In the pre-editing stage we enclose quotes of this kind in quotation-marks (" "), and the computer has instructions not to let the enclosed material through to the second processing stage. On the other hand, quotes from works not appearing in our archives, perhaps from no longer extant compositions, or quotes from works already processed but nonetheless interesting, because of their slightly altered

wording or because of the special antiquity or significance of the composition they appear in, are enclosed in a second kind of quotation marks (« »), and are passed along to Stage Two. All this refers to quotes in their entirety. If, however, quotes of the first kind discussed are interwoven with the wording of the text quoting them, in such a way that they form an integral part of its language, they are not enclosed in any quotation-marks, and so are let through by the computer to Stage Two.

(9) "THROUGHOUT THEIR GENERATIONS" (Ex. xxxi, 13). THIS LAW SHOULD OBTAIN THROUGHOUT THE GENERATIONS.

The first time the words THROUGHOUT and GENERATIONS appear, they are a Biblical quote; the second time, they form part of the language of the source (Midrash Mekilta).

This applies to phrases of foreign origin as well. Many of our sources are composed in Hebrew interspersed with words, phrases and even sentences in other languages, in particular - Aramaic. The Bible alrady exhibits verses and chapters in Aramaic. Both the Jerusalem and Babylonian Talmudim and some of the Midrashim are written in a mixture of Hebrew and Aramaic. Our rules provide that independent whole Aramaic phrases be discarded in the first stage of processing, the pre-edition of the text, and for this purpose we use the upper part of square brackets (⌜ ⌝). For example, a

Latin sentence within an English text:

(10) THE MODERN PHILOSOPHY BEGINS WITH THE ⌜COGITO, ERGO SUM⌝ OF DESCARTES.

If, however, the words of foreign origin are not an independent sentence, but rather form part of a Hebrew sentence, such as:

(11) HE IS THE ENFANT TERRIBLE OF MODERN ART.

We consider them part of the Hebrew language and refrain from using these special brackets.

Up to this point we described the symbols used in the first stage. Regarding the symbols we use in the second stage, similar ones serve our aims there as well. When we determine the lemma of an erased, added or corrected word, it is often unnecessary to add any signs at all in Stage Two, such as whenever the correction is evident from those signs added in Stage One, e.g., in (6) we determine lemmas for the words in square brackets, and the reader sees these and realizes that they refer to added words. However, in a case such as (3), we note:

Pre-edition: TO(0) Lemmatization: TO /[

i.e., the lemma represents the word after its correction.

If we are of the opinion that the word prior to its correction was also meaningful - and it is often the original, uncorrected, version which interests the scholar, for it may just reflect a word in use in ancient

(9)

times but no longer used later on - we note it thus:

Pre-edition TO(O) Lemmatization: TOO /(

i.e., before correction it was the word TOO.

This also applies to item (7):

Pre-edition: O(F)[R] Lemmatization: (1) OF /(

(2) OR /[

i.e., prior to its correction it read OF, afterwards - OR.

Concerning quotes and foreign phrases, if the excerpt is either entirely foreign or a complete quotation, we mark it with the appropriate signs and program the computer not to let it through to Stage Two. If, however, it is made up of words interlaced with others of the text, it is not marked at all in the first stage, but in the second the editor denotes it with single (unpaired) signs paralleling those used in Stage One.

In example (9), the first time they appear, the words THROUGHOUT and GENERATIONS are enclosed in quotation-marks and do not reach Stage Two. The second time, the same words are not marked with quotation-marks, and in Stage Two they are marked thus:

Pre-edition: THROUGHOUT Lemmatization: THROUGHOUT /"

GENERATIONS GENERATION /"

The non-English (Latin) words in example (10) are enclosed in ⌜ ⌝, and do not reach Stage Two, while the French words in example (11) reach Stage Two and are marked thus:

Pre-edition:	ENFANT	Lemmatization:	ENFANT	/⌜
	TERRIBLE		TERRIBLE	/⌜

These are typical examples. We use many other symbols as well; to denot lacunae in the manuscript, alternate readings, abbreviations and many others. I trust I have succeeded in presenting a general, but coherent picture of our system of symbols.

A Computer-assisted Analysis of a Hebrew Writer's Vocabulary

Dr. Reuven Merkin

The Academy of the Hebrew Language, Jerusalem

The Historical Dictionary Project of the Hebrew Language Academy processes literary and non-literary sources dating from different periods in the development of the Hebrew language, and prepares concordances and lexical and grammatical indices of each source or group of sources.

The Modern Hebrew Section, established ten years ago, deals with works written during the past two hundred and thirty years, wherever Hebrew composition took place. To date, one hundred and sixty sources, representing eleven different types of composition, have been selected for processing. These sources will provide over three million of the ten million excerpts the Project proposes to process, in order to provide a broad foundation on which to base a lexicographical description of Modern Hebrew.

This material, arranged according to fixed grammatical and lexical principles, will provide a solid, objective basis for whoever desires to describe the vocabulary of a period, of a literary type, or of an individual writer.

My research deals with the vocabulary used by the greatest Hebrew prose writer of the end of the nineteenth century, whose place in Hebrew literature may be compared to those of Dickens and Balzac in European literature.

Mendele is regarded as the creator of the modern prose styles of two literatures: Hebrew and Yiddish. His Hebrew composition bridges a period of sixty years (1856-1917). The first half of this period is noted for Mendele's publicistic writing and for his monumental zoological classic, "Natural History", which pioneered by supplying many animals with new Hebrew names. During the second half of the sixty-year period, it was his short stories and novels which won him fame.

Mendele's Hebrew style experienced significant changes from the first thirty years to the second. Biblical elements dominated his writing style during the earlier period, with considerable differences marking the various literary types. Later, the Biblical elements receded, while post-Biblical Hebrew became more dominant, especially "Rabbinic Hebrew", in the broadest meaning of this term.

Hebrew literature underwent a great change during the period of Mendele's literary activity. The years 1860-1890 are generally viewed - within the framework of modern Hebrew literary history - as marking the close of Haskala literature, which had first evolved some hundred years before, while the post-1890 years are considered "the Generation of Revival", the beginning of Modern Hebrew literature.

A parallel development took place in the Hebrew language itself, during Mendele's lifetime, most familiarly expressed by the reinstatement of Hebrew speech in Palestine towards the end of the nineteenth century. A new writing style was similarly created in East European Hebrew literature, a change generally identified with Mendele's stories, starting from 1886.

(3)

The researchers' impressions led them to the conclusion that Mendele's style changed drastically in 1886, with the publication of his first short story, "Hidden in Thunder". A quantitative analysis of Mendele's vocabulary exhibited by his various works shows that the change took place much earlier, more precisely between 1870 and 1875. The change did not occur simultaneously in all his writing styles, and its roots can be traced even earlier than 1870, but at that time the magnitude of the change and its direction were hardly predictable.

Mendele's varied literary activity during the earlier of the two periods in his career, together with the fact that he was active for sixty years, raise two questions concerning his Hebrew style:

(1) Do his different writing types - stories, articles, letters, and scientific material - exhibit differences in their vocabulary and in its grammatical aspects?

(2) Do his earlier compositions differ from his later ones in these same two respects?

The corpus selected to serve as the basis for an examination of these two questions includes eight works: four whole compositions, two anthologies of shorter items (articles and letters), and large extracts from the first two parts of his "Natural History". This corpus contains 313,554 word-appearances, which have been lexically processed and classified by the computer in accordance with the methods and principles of the Hebrew Language Historical Dictionary Project.

After its lexical processing, the corpus was tested quantitatively in many ways, questions such as the following being framed:

1. How lexically "rich" is the corpus, and each work it contains? That is, how does the number of word-appearances it contains compare with the corresponding number of lexical units?

2. How frequently do certain kinds of words appear, such as European-language loan words, proper nouns, and so on?

3. In how many different conjugations is each verbal root used? At the one extreme, we have the hundreds of roots used in a single conjugation only, while at the other, there are the few roots appearing in seven, eight or ten conjugations.

4. What is the actual frequency of each verbal conjugation, relative to others of its kind?

5. Where are the various verbal tenses and verbal noun forms (participles and gerunds) to be found?

6. How do nouns and verbs compare numerically?

7. How are the various noun patterns used?

8. How frequently do the different lexical items appear, from the thousands of items appearing once or twice in the entire corpus, to those encountered thousands of times?

9. Which lexical items - making up Mendele's "general" vocabulary - appear in all eight works composing the corpus, and which appear in only one or two of these works, being the "particular" vocabulary of that one specific composition or of one certain literary type?

The quantitative analysis of Mendele's vocabulary based on these questions - and many others not listed here - provide a positive, detailed solution to the problems of the linguistic differences between Mendele's various works, both generically and chronologically.

I now propose to describe one of the tests performed on Mendele's works, and to reveal its results.

Every word of the corpus, classified according to the principles of the Historical Dictionary Project as belonging to the verbal system, was listed by (1) its root; (2) its conjugation (there are ten of these); (3) its "tense" (past, future, imperative, participle, infinitive, gerund).

Our corpus - containing 313,554 word-appearances - includes 80,277 (25.6%) belonging to the verbal system, derived from 1727 different roots (see Table 1). At most, half (48.6%) of these appears in only a single conjugation. Another 41.4% of the roots appear in two or three conjugations each - together, 90%. The remainder (10%) appears in four conjugations or more.

Regarding each conjugation derived from a given root as a separate entity, the 1727 roots provide 3319 verbs*, i.e., each root appears on the average in just under two conjugations (more precisely, in 1.92 conjugations). The verbs in our corpus have an average frequency of 24 (80,277 : 3,319=24.17), in contrast to the nouns, the average frequency of which is 41.

---

*36.85% of 9,011 lemmata found in the corpus.

Similarly, considering each "tense" of a given verb an independent unit, the 3319 verbs appear in 7800 "tense" units, i.e., each verb appears on the average in 2.35 "tenses".

Verbs derived from single-conjugation roots appear on the average in two "tenses" (in contrast to the general average of 2.35), their average frequency is only 10.5 (as compared with the general average of 24), and the average frequency of each of their "tenses" is 5.2 (contrasting with the general average of 10.3).

In comparison, verbs derived from roots appearing in five conjugations or more appear in an average number of "tenses" approaching 3, their average frequency is 40, and that of each of their "tenses" - 14.6.

In other words, the most common verbs in the corpus are derived from roots appearing in many conjugations. It is however just possible that historically the causal relationship is to be inverted; the commonest verbs in the language having engendered a situation where their roots appear in many conjugations.

Another conclusion may be drawn concerning the various works making up the corpus: an examination of the frequencies of the different verbal conjugations in the different compositions shows that certain conjugations are more "stylistically characteristic" than others, meaning that stylistic similarity between two works may also be measured by their conjugation frequencies.

(7)

Modern Hebrew linguistics views the verbal conjugations (not including the specifically passive ones) as independent lexical and semantic units, while the root is regarded as an abstraction (see Table 2). The results of the examination of our corpus suggest that the root may perhaps best be considered - at least historically - a semantic and lexical unit, while the verbal conjugations seem to be mere flexions.

Table 1. DISTRIBUTION OF VERBS

number of conjugations	number of verb radicals found	number of verbs*	average number of tenses per verb	average per verb	occurrence per tense
1	839	839	2.01	10.5	5.2
2	468	936	2.31	23.6	10.2
3	247	741	2.43	26.3	10.8
4	96	384	2.57	33.2	12.9
5	54	270	2.80	40.9	14.6
6	16	96	2.65	34.4	13.0
7-10	7	53	2.97	51.3	17.3
TOTAL	1727	3319	2.35	24.2	10.3

Table 2. EXAMPLES OF VERB RADICALS IN DIFFERENT CONJUGATIONS

the radical K-T-V

the conjugation	the verb	
pa'al	KaTaV	"write"
nif'al	niKhTaV	"be written"
hitpa'el	hitKaTteV	"exchange letters"
hif'il	hiKhTiV	"dictate"
huf'al	huKhTaV	"be dictated"

the radical K-N-S

nif'al	niKhNaS	"enter, go in"
pi'el	KiNneS	"bring together"
pu'al	KuNnaS	"be brought together"
hitpa'el	hitKaNneS	"come together"
hif'il	hiKhNiS	"bring in"
huf'al	huKhNaS	"be brought in"

Computerized Lemmatization of Non-vocalized Hebrew Texts

Zion Busharia

The Historical Dictionary of the Hebrew Language
The Academy of the Hebrew Language, Jerusalem

One of the stages in the amassing of material for the Historical Dictionary is the classification of preprocessed textual quotations according to their lemmata, and the preparation of concordances. An alphabetic sorting of Hebrew word-tokens is not particularly useful in determining these lemmata for three reasons:

1) The morphological structure of Hebrew being a language of roots and flexions, which give rise to a vast number of forms belonging to a single lexical item. These forms do not appear in one alphabetically consecutive group, because the item, the kernel of the word (and of its accepted dictionary form), is merely a core to which various elements are prefixed, infixed and suffixed. These elements divide up into four classes:

a. Lexical particles: words of a single letter never appearing independently, but rather always prefixed to the following word, whether noun or verb.

b. (Verbal) prefixes: letters denoting conjugation and tense.

c. (Both nominal and verbal) suffixes: determining status, person, gender and number.

d. (Both nominal and verbal) pronominal elements: denoting possession (with nominal forms) or the objective case (with verbs).

For example, the dictionary entry (and accepted lexical form) of the verb לִמַּד LMD (to teach) is composed of the three letters LMD (most Hebrew roots being triliteral). Person is expressed in the past tense - I taught, you taught, etc. - by adding suffixes to these three leters (class c, above): LMDTY, LMDT, etc. Prefixes and suffixes are added to the root-letters to express the various persons in the future tense (classes c,d, above): they will teach, you (f. sing.) will teach, etc. - YLMDW, TLMDY, etc. Where the desired expression is "they will teach him", "you (f. sing.) will teach them", etc., pronominal elements are added to the previously-

affixed prefixes and suffixes: YLMDWHW, TLMDYM, etc. There are a number of one-letter words which are prefixed to the word, preceding any other prefixed elements, to express and, that, where, to, from, in, etc.; e.g."and when they (will) teach him": WKŠYLMDWHW. This word has seven letters over and above the three-letter root, four of which precede the root-letters and thus complicate the problems of alphabetical order. Moreover, this root is a strong one, with all its root-letters appearing in each of its inflected forms. But Hebrew has many weak roots as well, which undergo such morphophonetic changes in declension that only two - and occasionally only one - of the original root-letters appear in several of the inflected forms, surrounded by a large number of prefixes, suffixes etc.

2) The second factor complicating the determination of Hebrew lemmata is the lack of a clear system of vowel-notation in Hebrew orthography, which is basically consonantal (as in the other semitic languages). This lack results in a large number of homographic lemmata and forms.

With the passage of time, a few letters - especially י y and ו w - began to serve as vowel-signs, but the number of homographs is only slightly reduced in consequence, since each of these letters represents two different vowels (waw - o,u; yod - i,e), in addition to the consonants they continue to represent. Moreover, Hebrew exhibits two consonants, ś - š, denoted by one single letter, and distinguished in vocalized spelling by a diacritical point, but indistinguishable in the usual non-vocalized orthography.

3) Furthermore, it should be recalled that Hebrew orthography in general, and particularly insofar as the use of w and y as vowel-markers is concerned, is not at all uniform, neither today nor, especially, in ancient days, so that a single word may be spelled in several ways, and sometimes quite a few.

Here are a number of examples to clarify the matter of numerous homographs in the language:

Homographic lemmata

In a Hebrew dictionary the letters (דבר DBR) appear under no fewer than 11 different lexical lemmata:

A.

Verbs

1. DBR	DiBbeR	he spoke
2. DBR	DuBbaR	(it) was spoken
3. DBR	DoBeR	speaking (part.)
4. DBR	DaBbeR	speak ! (impv.)
5. DBR	DaBbeR	(to) speak (inf.)
6. DBR	DaBuR	spoken (part. pass.)

Nouns

7. DBR	DaBaR	a thing
8. DBR	DaBbaR	(a) a leader (b) a raftsman
9. DBR	DiBbeR	a statement, command
10. DBR	DoBƐR	a cattle-pasture
11. DBR	DƐBƐR	a plague

Morphological homographs

The letters (מדבר)MDBR can be interpreted in 12 different ways:

B. MDBR

1-6 M(from)+DBR	see A,5,7-11	
7. MDBR	MəDaBbeR	speaking (part.)
8. MDBR	MəDuBbaR	spoken (part. pass.)
9. MDBR	MuDBaR	exterminated (part.)
10. MDBR	MiDdaBbeR	speaking (to himself) (part.)
11. MDBR	MiDBaR	(a) a desert (b) a voiced expression
12. MDBR	MaDBeR	an instrument for the elimination of various disturbances

Words variously divisible

Initial or final letters may be either parts of the root or affixes:

C. 1. K-LB/KLB KəLeB/ KƐLƐB as a heart/dog

2. ŠM-N/ŠMN/Š-MN ŠəMaN/ŠƐMƐN/ŠƐMiN their name(pl.f.)/oil/that (is)from

Hebrew exhibits a system of vowel-signs, generally written under the letters. Such a vocalized text would really solve a considerable portion of the homographic problem; however, these vowel-signs appear in only a minority of Hebrew texts, mainly Bible texts, poetry and children's literature; whereas most Hebrew texts are non-vocalized. The problem of homographs thus remains the most difficult problem of computer-assisted lemmatization.

Our examination of numerous methods of automatic computerized lemmatization has led us to the following conclusions:

a) A computer may be used to determine lexical items, but no automatic system can relieve the lexicographer of the responsibility of double-checking the computer's determination, for the Dictionary requires a precise determination of the lemma corresponding to the word-token in context.

b) No system can ever supply the memory of the computer with all possible language forms, since each new text may well introduce previously unknown forms and lexical items, such as proper nouns, loanwords of foreign extraction, abbreviated forms, initialled abbreviations, and so on.

c) The optimal system is one which makes minimal demands of the lexicographer and entails maximal exploitation of the computer.

These considerations led us to reject the mechanicographization of the morphology, i.e., the committing of the grammar and the lexicon to the memory of the computer. This system would require enlarged memory-banks, and a slower search by the computer, and would result in a number of lemmata being offered for each word - an extra burden for the lexicographer to shoulder.

Our lemmatization programs entail an indeterminate form-bank, compiled from the very start on the basis of forms already appearing in the processed texts (a "form" being defined here as a string of letters between spaces), and the corresponding items listed by the lexicographer. The very first text was analyzed manually, and yielded the primary form-bank, which was then applied to the second text; the cumputer produced lemmata for some of the forms

of this second text on the basis of the primary form-bank, and for those forms for which no such lemmata were provided, lemmata were determined manually by the lexicographer. These formed a supplement to the form-bank, which was then applied to the next text, and so on. In this way, the form-bank keeps continuously growing.

The form-bank is made up of two kinds of records: a- a form record embracing the form itself, its corresponding lemma(ta) and the frequencies of these lemmata in the texts already processed; b- lemma-record including all the forms appearing under that lemma. The lemma-records serve to update the form-bank and for information retrieval.

The lemmatization process is as follows: the computer picks out a word (a form) from the text. It then removes all editing-signs (e.g. question-marks, parentheses, etc.) if there are any, according to predetermined rules. The computer also removes the particle ו (and) from its word-initial position (a very common Hebrew prefix appearing in combination with a considerable percentage of words), and then searches the form-bank for this form taken from the text. If it is to be found there, together with the lemmata listed for it in earlier stages, the computer lists these items as lemmata suggested for the form from the text. This is to say, the lemmata suggested by the computer are not all the theoretically-possible lemmata, according to Hebrew grammar and lexicons, but only those already listed for this form in texts previously processed by our Project, i.e., lemmata already encountered empirically.

As for forms in the text which do not appear in the form-bank , the computer does not suggest lemmata for these, and it is the lexicographer who adds their lemmata manually. This information augments the form-bank and brings it up-to-date.

This system results in the constant enlarging of the form-bank, as more and more texts are processed. This enlargement of the form-bank brings about a decrease in the percentage of words encountered in subsequent texts for which the computer has no lemmata to offer, and on the other hand, an increase in the number of multivalent forms, i.e. forms which have more than one lemma.

Several criteria have been adopted to assist the lexicographer in selecting the appropriate lemma for each word, and to render the entire process more efficient:

a) The Statistical Criterion:

1. The lemmata suggested for a given form are arranged according to their frequency in the texts already processed. This frequency is brought up-to-date with the processing of each additional text. The most common lemma is listed first, and if it suits the given context, the lexicographer adds nothing, and this item is treated as the correct one (the default).

2. If several lemmata are available for a given form, one is of a relatively high frequency, and the others of low frequencies; if these low frequencies are less than 1 %, the computer does not propose them at all; likewise, if their frequencies are less than 7 % , and the absolute number of their occurences is under 10.

3. The computer never produces more than seven lemmata for any one form.

4. The significance of the frequency of a proper-noun lemma is reduced by 80 %.

These rules are changed empirically to determine the optimal parameters.

b) The Historical Criterion:

Since the vocabulary and orthographical systems change from one period to another, we compile separate form-banks for each period processed by our Project. One general form-bank may be used, but in that case the relevant period is to be noted in the form-record.

At present we are processing two periods: ancient literature (200 B.C.E. - 1000 C.E.) and modern literature (1750 C.E. to the present day). Each period has its own form-bank and its own criteria for selecting lexical items.

c) The Literary Criterion:

We do not apply this criterion in our Project as yet, but when we have different kinds of processed texts we plan to note the literary type of the texts in the form-bank records, such as rabbinical literature, scientific literature, newspapers, and so on.

We have been using this lemmatization system for over ten years, with very good results. The ancient literary texts most recently processed yielded the following results:

The computer suggested lexical lemmata for over 97 % of the words in the text.

It proposed a single lemma for 75 % of these words.

The default was good for 95 % of the words for which the computer suggested lemmata (i.e., no other effort was necessary, such as listing an item, vocalization, verification, etc.). The deviation is ±2 %, depending on the kind of text.

The lexicographer had to intervene in less than 9 % of all the words in the text.

The forms for which the computer is unable to suggest lexical items are, generally speaking, infrequent. In the Tosefta, Tractate Zera'im, for example (ancient literature), the average frequency of a form is 6, whereas that of a form not found in the form-bank is only 1.5. In modern literature, we found an inaccuracy in this statement, as a number of quite common forms were encountered, with no previous listings in the form-bank. An examination of the details revealed that these are generally proper-nouns, place names, or fashionable foreign words, appearing in texts dealing with contemporary problems .

After the processing of Tannaitic-period texts was completed, covering some 830,000 words, we carried out a number of statistical experiments to verify the percentage of homographic forms in these texts, and we found that the number of multivalent forms is only 8.5 %, but the percentage of the tokens of these forms is over 30 %, in other words: many multivalent forms are extremely common. For example: of all form-occurences in this corpus of a frequency of 1-4, 20 % are multivalent; of those whose frequency is 5 or more, 50 % are multivalent.

Today, after processing over a million words in ancient literary texts, the form-bank contains 63,000 forms in all. Forms beginning with ו w (the conjunctive particle "and", very common in Hebrew) have not been included.

(8)

We are trying to perfect our lemmatization programs, mainly by removing all prefixed particles from the forms, as we did successfully in the case of ו w. The remaining particles are less frequent than w, and the attempt to remove them from the forms has raised by about 3 % the number of words for which the computer could suggest lemmata, while slightly lengthening the time taken by the computer to search its memory-banks for them. This step towards perfection is still being checked.